图书情报工作
第 65 卷 第 16 期 2021 年 8 月

基于 LDA-BERT 融合模型的弱信号识别研究*

■ 杨波[1,2] 邵婉婷[1,2]

[1] 江西财经大学信息管理学院 南昌 330013 [2]江西财经大学信息资源管理研究所 南昌 330013

摘 要：［目的/意义］针对现有弱信号全自动识别研究尚不完善的问题，提出基于 LDA-BERT 融合模型的弱信号全自动识别方法。［方法/过程］基于无监督的 LDA 主题模型对文本数据集进行主题分类，构建主题和术语双层过滤函数从主题分类的结果中提取早期预警信号，通过紧密中心度、主题权重以及主题自相关性三大度量函数评价主题的弱性，并基于主题内术语的归一化频率和概率提取出弱信号。最后，运用 BERT 深度学习模型从语义层面对弱信号上下文及其类似词进行扩展。［结果/结论］以 2021 年 1 月初疫情重爆发事件为例，使用爆发前三月的社交媒体新闻数据集对构建的系统模型进行验证。实验结果表明，该方法可有效检测出相关弱信号，并挖掘出弱信号随时间推移逐渐增强的演化特性。此外，该融合模型在实现弱信号全自动识别的同时，也表现出较单一模型更强的结果可解释能力。

关键词：弱信号 LDA-BERT 融合模型 新冠肺炎疫情

分类号：G250

DOI：10.13266/j.issn.0252－3116.2021.16.011

1 引言

在大数据时代的背景下，人们的决策行为更多依赖于所获取数据及信息的分析，而不是仅凭借直觉和经验，竞争情报作为构建数据及信息分析的基础，其获取、收集与识别工作至关重要[1]。弱信号是竞争情报前瞻性研究中的重要组成部分，为决策者预测未来的机会与风险提供有益参考。弱信号与大多数信息一致，都是从大量的数据中提取而出，通过合理的推断与联系，形成对人类有价值的信息，但由于其具有预见性的特点，也被称为预警信号[2]，忽视弱信号就是轻视甚至压制可能阻止错误决策的警示信号，如同驾车闯红灯，定会导致失败[3]。因此，弱信号的识别研究对决策者及时感知市场的机遇与威胁，制定利于长远发展的管理决策有一定的现实意义。

目前，识别弱信号并预测未来情况已成为许多研究人员的目标，许多技术用于从词或文档中获得最大洞察力，但大多需要人类专家的协助检测。传统的主题建模技术显示了其完全自动化的能力[4]，因此，本研究使用一种广为人知的主题模型，即潜在狄利克雷分布（Latent Dirichlet Allocation，LDA）。LDA 是一种无监督的机器学习技术，可根据输入的文档集及指定的主题数来独立运行，不需要手工标注的训练集，在弱信号识别过程中可实现全自动地从数据集中抽取出主题及主题所对应的关键词。

而 LDA 主题模型的提取结果并非都为弱信号，仍存在具有明确分类指向的强信号和无法揭示出具体含义的噪声信号[5]，因此，为检测出隐藏、重要且被限定为弱信号的单词还需进一步对其过滤。同时，由于弱信号具有稀有、微量的特点，导致提取出的弱信号数量较少[6－10]。为充分地对弱信号进行自动化检测，本研究提出一种基于 LDA-BERT 融合模型的弱信号全自动识别方法，通过构建主题和术语双层过滤函数，从 LDA 主题分类的结果中提取早期预警信号，通过紧密中心度、主题权重以及主题自相关性三大度量函数对主题进行过滤，并基于主题内术语的归一化频率和概率提取出弱信号。最后，为弥补 LDA 词袋模型的不足，增强模型结果的可解释性，本文运用 BERT 方法对每个

* 本文系国家自然科学基金项目“基于免疫方法的新创企业成长风险管理知识服务模型研究”（项目编号：72064015）和江西省社会科学规划重点项目“面向新创企业成长风险管理的知识服务机制研究”（项目编号：19TQ01）研究成果之一。

作者简介：杨波（ORCID：0000－0001－6012－9007），副教授，博士，博士生导师，E-mail：yangbo@jxufe.edu.cn；邵婉婷（ORCID：0000－0002－0700－0113），硕士研究生。

收稿日期：2021－04－11 修回日期：2021－06－11 本文起止页码：98－107 本文责任编辑：易飞

过滤后的主题文档进行上下文的预测,以获得更多与弱信号语义相关的单词。将该模型应用于社交媒体新闻数据集,以检测 2021 年 1 月初疫情重爆发的相关弱信号。

2 相关研究

"弱信号"最早由 H. Igor Ansoff 在 1975 年提出,并将其定义为"未来可能发生变化的症状"[11]。他认为弱信号是对外部或内部的警告,这些警告具有不完整性,无法准确预估其影响,一个组织要及时应对不确定的环境,就必须提前做好准备,对可能蕴含威胁和机会的信息迹象作出反应[11-12]。此后,B. Coffman、P. Rossel、S. Mendona 等学者对弱信号的概念做出了进一步的补充,他们认为弱信号具有以下特征:不易追踪,与夹杂的噪声难以区分;琐碎、易被忽视,却对未来可能造成重大影响;未来改变和趋势的早期线索[6-10]。

我国弱信号的相关研究起步较晚,但也提出了相对深刻的见解。沈固朝认为弱信号是通过对组织竞争环境中迹象的观察、业内人员意见的分析,对未来的趋势波动做出早期判断[13]。单彬总结出弱信号"弱"的四大原因:①能被感知的弱信号量较少;②有效的信息难以被捕获;③误导或虚假信号与有效信息并存;④收集信号的成本和精力有限[14]。赵小康指出弱信号具有在生长过程中表现渐趋明显、确定性不断增加、包含的有效信息量逐步丰富、作为决策依据的情报价值持续提高 4 项主要特征[15]。

目前,弱信号的识别过程缺乏自动化,大多研究依赖于手工输入或专家意见[4]。如 I. Griol-Barres 等利用科学、新闻和社会来源的异构和非结构化信息对弱信号进行定量检测,应用多词共现分析法对人工挑选的部分关键词进行分析,并通过自然语言处理提取准确的结果[16-17]。J. Yoon 提出一种在专家给定关键字的前提下,基于文本挖掘的弱信号主题识别方法,并通过太阳能电池相关的网络新闻报道,说明了该方法的可行性[6]。邓胜利等通过专家给定系数下的层次分析法和隶属度函数对弱信号进行定量识别[18]。这些方法需要大量的人工耗费,且使得弱信号识别的结果具有较强的主观特性。

与此同时,学者们也着力于运用诸如深度学习和神经网络之类的技术来充分对互联网上不断增加的文本数据进行预见性分析。自然语言处理技术(NLP)能够很好地从文本数据中提取见解[19],其中单词嵌入技术能精准地捕获词语之间的相似性和基于上下文预测单词[20]。B. Dieng Adji 等提出一种嵌入式主题模型,该模型将常规主题模型与单词嵌入结合在一起。但是,与未标记的数据相比,这些技术在应用于标记的数据时可提供更好的结果[20-21]。而在 Web 文章中检测弱信号的情况下,文本数据通常没有标签。因此,基于深度学习的 NLP 技术不能确保弱信号检测过程的完全自动化。

弱信号检测的全自动化研究尚处于起步阶段,相关的论文和项目数量较少,在全自动化识别过程中主题模型被广泛应用于隐藏信息的检测。如 L. Pépin 使用动态 LDA 检测弱信号,即对不同时间下的文本使用 LDA 算法提取主题,并使用主题演化的可视化散点图来检测弱信号[22]。T. Gutsche[23]提出一种运用动态主题建模和时间序列分析对弱信号进行自动检测和预测的方法,取得了较好效果。本研究遵循与其相同的完全自动化方法,选用 LDA 从社交媒体新闻数据集中提取主题及主题所对应的关键词信息。而 LDA 主题模型的提取结果除了弱信号外仍存在强信号和噪声信号[5],因此,还需对 LDA 提取结果进一步过滤。此外,庄穆妮等指出 LDA 词袋模型的不足,即在 LDA 中一篇文档仅为一组单词的集合,词与词没有先后顺序,无法很好地结合上下文信息[24]。针对此问题,J. Maitre 等提出运用 Word2Vec 方法增强 LDA 主题模型[25],以获得更多类似弱信号的单词。但 L. Kahyun 等在比较 NLP 领域中 Word2Vec 和 BERT 算法时,发现后者更能体现词语在语义和语法方面的复杂性,对解决一词多意的问题更有帮助[26],即在 LDA 模型增强中表现更优异。

综上所述,目前关于弱信号识别的方法存在各自的局限性,主要表现在提取与识别过程中多依赖人类专家的协助,对于弱信号全自动识别的方法研究尚不完善,且提取结果数量较少,难以挖掘其之间关联性,导致可解释能力不高,预警效果并不十分理想。因此,为实现全自动弱信号检测,弥补单一 LDA 词袋模型的不足,增强识别模型结果的可解释性,本研究将引入 BERT 模型对 LDA 的提取结果作进一步处理分析,构建 LDA-BERT 融合模型,在对主题和术语进行双层深度过滤的同时也对提取出的弱信号进行语义上的扩展,以获得更好的弱信号识别效果。

3 弱信号自动识别方法框架

3.1 方法概述

为减少人类专家的干预,设计一个全自动弱信号

识别方法，本研究考虑使用与主题建模相关的无监督文本挖掘技术。其中，LDA 常用于从文本数据集中提取趋势主题。与依赖关键词进行弱信号检测[6]的研究相比，主题模型更多地是考虑单词代表的意义，而不是其本身。本文运用 LDA 主题模型寻找可能导致弱信号的主题，但不接受所有主题中都含有弱信号，也不认为主题中的所有术语都为弱信号，即除弱信号外仍存在具有明确分类指向的强信号和无法揭示出具体含义的噪声信号[5]。因此，本文提出主题过滤和术语过滤双层过滤模型，用于仅提取潜在的弱信号。其中，主题过滤模型基于紧密中心度、主题权重以及主题自相关性构建主题弱性评价函数，并基于此函数值提取可能包含弱信号的主题。术语过滤模型用于从主题过滤提取的可能包含弱信号的主题中进一步提取弱信号相关术语，主要依赖于主题内术语的归一化频率和概率判断是否为弱信号。但由于弱信号具有稀少、微量的特点，导致提取出的弱信号数量较少，难以发掘其之间存在的关联性，使模型的可解释能力不高。为解决此问题，参考 J. Maitre 等运用 Word2Vec 模型增强提取结果[25]的方法，并采用在语义和语法方面表现更优异的 BERT 深度学习模型对弱信号从上下文综合进行语义扩展，以获得更多的可解释性预警信息。

该方法的框架如下：①收集数据，本研究收集了一段时间的社交媒体新闻内容作为弱信号识别研究的输入。②弱信号识别，包括数据预处理和弱信号过滤两部分。数据预处理是对收集的文本集进行去停用词、分词操作。弱信号过滤包括运用 LDA 主题模型识别主题、对提取出的主题和术语过滤，以寻找潜在的弱主题和弱信号。③弱信号输出，运用 BERT 模型词嵌入来增强识别出的弱信号并输出。如图 1 所示：



图 1 弱信号自动识别过程

3.2 数据收集和预处理

弱信号识别任务中，文本数据集的质量与弱信号检测结果的准确性、预见性有直接的关联，本研究运用 python 工具进行数据收集和预处理工作，基本步骤如下：

（1）文本数据收集。运用网络爬虫技术，从互联网中收集一段时间的新闻数据。本研究以社交媒体新闻为研究对象，因其具有传播范围广、传播及时性强、传播速度快等特点，对弱信号识别而言是较优的数据源。

（2）文本集清洗与分词。对收集的新闻数据集进行基于中文停用词表的清洗，目的是过滤其中不相关、无意义以及非文本的信息。并运用 jieba 对清洗后的数据进行分词，最终得到可用于系统输入的数据集。

3.3 基于 LDA-BERT 融合模型的弱信号自动识别

3.3.1 LDA 主题模型训练

LDA 主题模型又称为隐含狄利克雷分布，是在预先规定的主题数量下通过最大化词语共现的概率从文本集中查找潜在和隐藏的信息，如在一篇新闻报道中“足球”“运动”之类的词总是同时出现，即可把其归为体育类。D. M. Blei 等认为 LDA 能很好地对文档主题进行抽取[27]。LDA 主题模型的主要挑战之一是确定最优的主题数 k，主题数过多会导致主题分布不够集中，主题之间相似性较高，主题数过少则导致主题的内容过于宽泛，没有明确的分类指向[5]。超参数 α 和 β 的值分别表示文档主题密度和单词主题密度，它们在建立主题和术语之间的一致性上发挥着重要作用。

目前，研究人员提出确定最佳主题数 k 的主流方法有困惑度法和一致性法。困惑度值越小，则主题分类的结果越优，但赵凯等学者在进行主题分类时发现随着主题数量的增加，其模型困惑度值逐渐递减，难以确认最佳主题数 k[28]。与此同时，黄佳佳等学者提出用一致性法来权衡主题质量[28-31]，并发现基于此提取出的主题具有较高的可解释性，因此本研究遵循这种方法，并应用[32]提出的主题相关性度量值 c_v 来确定最佳主题数。

为了找到一致性最高的模型，本研究采用控制变量法进行测试，每次运行仅改变主题数 k 的值，并保持其他参数值不变。使用 c_v 值作为一致性度量，并基于滑动窗口、标准化点互信息（NPMI）和余弦相似度确定其值，然后返回一致性度量最高的主题数 k 作为模型的最优结果。

3.3.2 主题过滤

本节中提出的主题过滤函数，有助于评估主题含有弱信号的可能性，并用于对 LDA 主题模型提取出的主题进行过滤，该方法由 Logistic 函数推导而出。Logistic 函数常用于说明人口的进步和增长，但在语言学中被用来模拟语言变化，一个边缘的术语随着时间的推移其传播速度会增加，但如果它是弱信号，传播速度

增加后将仍处于边缘状态[33]。基于此,本研究将从主题自身分布特性及主题发展特性两方面着手,创建主题弱性评价函数,挖掘自身表现弱的主题,如表现出与其他主题关联不紧密、在主题分布中占比不高等特点,且在发展中长时间处于边缘状态的主题。定义如下三大度量函数函数以确定主题的弱性:紧密中心度、主题权重以及主题自相关函数。

(1)紧密中心度通过主题与主题之间的距离表示其相似性。许多距离度量可以用来计算相似性,如 Jaccard 距离、余弦距离和 Hellinger 距离。L. Pépin 等学者发现当距离测量呈现出 S 形变化时,能最有效地表示文本之间的相似度[22]。基于此原则,本文选用 Hellinger 距离计算主题 z 的紧密中心度 $CC(z)$,其中,h 表示 Hellinger 距离:

$$CC(z) = \frac{1}{\sum_i h(z, z_i)} \qquad 式(1)$$

(2)主题权重模型内相关主题的一致性代表着主题的意义分配。因此,本文基于主题 z 的一致性和所有主题一致性的总和来定义主题 z 的权重值 W,其中,$Coh(z)$ 表示主题 z 的一致性大小:

$$W(z) = \frac{Coh(z)}{\sum Coh} \qquad 式(2)$$

(3)自相关。自相关性是目前盛行的数据趋势分析工具,趋势分析是基于以往数据对未来可能发生情况的推测,它量化并解释了随着时间的推移混乱数据中的趋势和模式。自相关描述了同一变量在不同时期之间的关系,即变量对应值与其滞后变量对应值线性相关。而在新闻数据集中,与某个主题相关的文档频率会随着时间而改变,因此每个主题在几天内的自相关性可帮助过滤出可能不包含弱信号的主题。自相关函数 AC 定义如下,其中 $Cov(z)_k$ 是主题 z 滞后 k 期的协方差,$Var(z)$ 是主题 z 的方差:

$$AC(z) = \frac{Cov(z)_k}{Var(z)} \qquad 式(3)$$

利用上述 3 个度量函数组成评判主题弱性的函数 WK,其函数值越低,主题内含有的术语越弱,但当其足够低时也可定义为噪声。定义主题 z 的弱函数 如下:

$$WK(z) = \frac{W(z) \times CC(z)}{1 + \exp^{-(AC(z))}} \qquad 式(4)$$

根据弱信号的定义,稀有是其主要特征,且随着时间的推移,它们的运动是缓慢的。因此只有 WK 函数低值对应的主题才被认定为弱主题。根据帕累托原则,弱信号形成的信息不超过 20%,且人类专家将噪声的阈值范围定义为0%至2%[34],表示文本中携带无意义信息单词的概率。基于此,本文决定忽略 WK 函数的低值情况,并定义新的筛选阈值:噪声低于 1%,弱信号低于 15%。文本中的信号分布情况如图 2 所示:

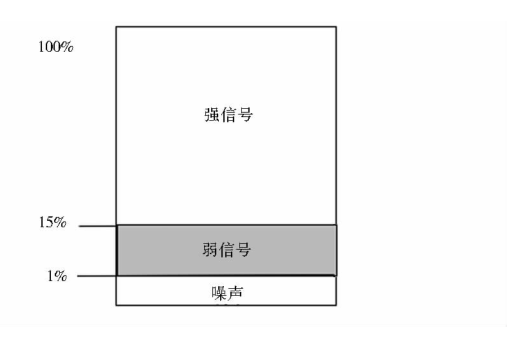


图 2　文本信号分布

3.3.3　术语过滤

基于定义的主题过滤函数能提取出可能包含弱信号的主题,但这些主题内的术语不一定都为弱信号,因此本节将探讨如何从这些术语中有效地提取弱信号。

J. Chuang 提出独特性和显著性两种术语衡量标准来判断某一主题中术语所传达的信息,以获得可理解的主题[35]。其研究发现单词由潜在主题生成的可能性与主题的边际概率之间的差异产生了显著性,而该显著性是属于的总体频率和独特性的产物。同时,C. Sievert 等通过主题内不同术语的相关性以寻求该主题内最相关的术语[36],并取得相比于概率指标更优的结果。

综合上述,基于术语在主题中的概率和术语与主题之间的相关性,本研究提出一种新的术语过滤函数 $PW(w)$,其中,$NF(w)$ 是主题 z 中术语 w 的归一化频率,$\varphi(w)$ 表示主题 w 中术语 的概率。

$$PW(w) = \frac{NF(w)}{1 + \exp^{-(\varphi(w) \times \log(\varphi(w)))}} \qquad 式(5)$$

同时,基于 3.3.2 主题过滤中所述,弱信号具有稀有性,因此本文仅提取 PW 函数值在 1% 至 15% 的项。

3.3.4　弱信号输出

在主题过滤和术语过滤两层过滤函数下,能很好地对弱信号进行识别与提取,此外对结果的分析与理解也至关重要。而弱信号稀有、微量的特点导致提取出的弱信号较少,为进一步获得与所提取弱信号相关的单词,提高模型结果的可解释性,本研究使用 BERT 深度学习模型增强弱信号提取结果。BERT(双向 Transformer 编码表达)模型由谷歌 2018 年推出,以 Transformer 算法为主要框架,能更好地捕获语句中的双向关系,并使用遮蔽语言模型 MLM(Mask Lan-guage

Model）和句子预测 NSP（Next Sentence Prediction）的多任务训练目标，使模型的结果达到了全新的高度[37]。其中 BERT 的模型结构如图 3 所示：

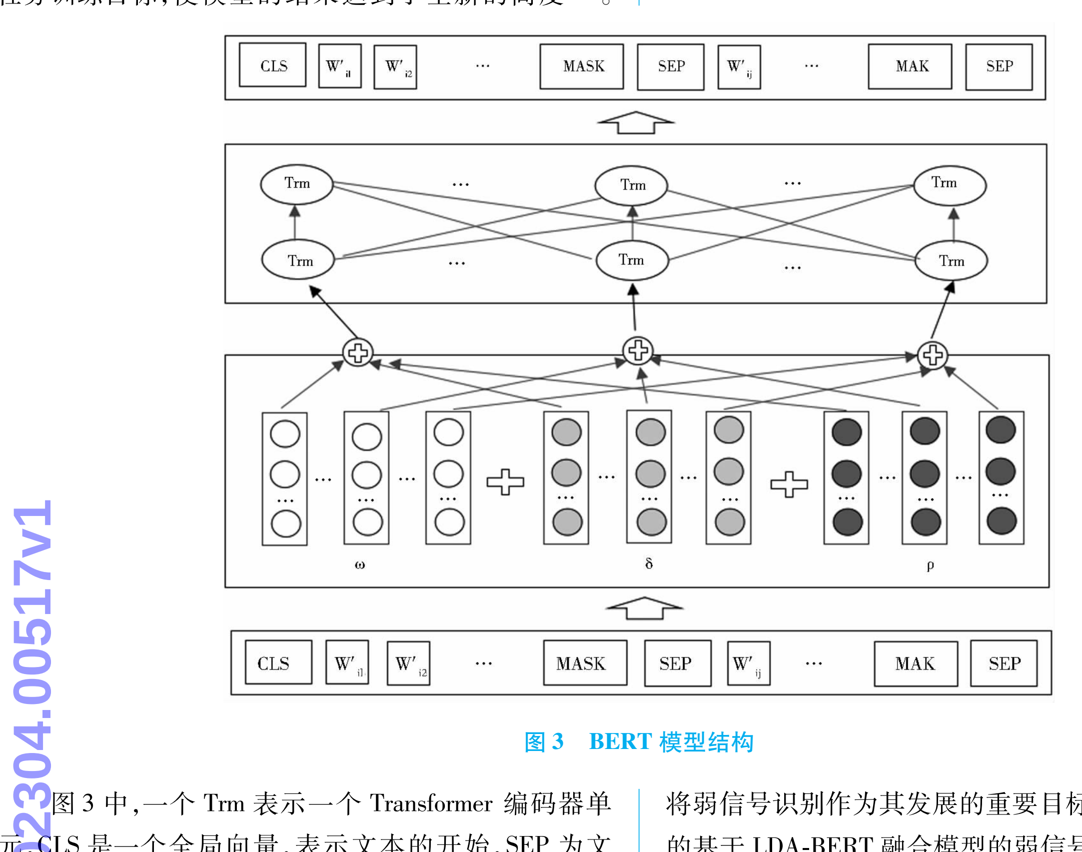


图 3　BERT 模型结构

图 3 中，一个 Trm 表示一个 Transformer 编码器单元，CLS 是一个全局向量，表示文本的开始，SEP 为文本连接符，MASK 为遮盖字。将提取出的每个弱信号单词看为一个向量，重建单词上下文，使语料库中共享公共上下文的单词在语义空间上相互接近，并扩展与提取结果相似的弱信号。将弱信号 $d_i\{w_{ij}|j\in\{1,2,\cdots,m\}\}$ 输入 BERT 模型后，首先在词向量（ω）、文本向量（δ）和位置向量（ρ）3 个维度上将弱信号中的每个字向量化为 $w_{ij}(\omega+\delta+\rho)$，再传入双向 Transformer 编码器中，最后输出融合全文本语义信息的向量集 $d'_i=\{w'_{ij}|j\in\{1,2,\cdots,m\}\}$。

该方法类似于 Word2Vec 中的 Skip-gram 模型，即根据当前单词来预测其上下文信息，但不同的是 BERT 方法较 Word2Vec 方法更多地从语义和语法两方面进行考量，且具有更丰富、完善的语料库，使其在单词语义扩展上表现更优异。本文遵循以往学者的研究[24,26,37]，运用基于 Google 预训练集的 Fine-tunning，将每个过滤的弱信号作为 BERT 模型输入，在经过训练后输出与提取弱信号高度相似的单词列表，以突出基于新闻数据集提取的弱信号及增强弱信号之间的关联性，获得更强的模型可解释能力。

4　实证研究

弱信号在竞争情报中占有重要地位，多数企业也将弱信号识别作为其发展的重要目标。本研究将提出的基于 LDA-BERT 融合模型的弱信号自动识别方法应用于微博等社交媒体发表的网络新闻，以检测 2021 年 1 月初疫情重爆发的早期预警信息。通过网络爬虫工具收集 2020 年 11 月 1 日至 2021 年 1 月 10 日的社交媒体新闻数据共计 14 486 篇，并运用 Python 开源库 jieba、Gensim 等对其进行分词、主题建模和自然语言处理等操作。

4.1　LDA 主题模型训练结果分析

为找到最优主题模型对应的主题数 k，本研究运用 Gensim 库中的 LdaModel 模块和 pyLDAvis 可视化工具，通过计算不同主题数下的一致性度量 c_v 值及其主题分布情况进行综合评判。

首先，本文对已进行清洗、分词等预处理操作的 2020 年 11 月 1 日至 2021 年 1 月 10 日的社交媒体新闻数据集进行 LDA 主题建模。其次，运用控制变量法测量不同主题数 k 下的一致性度量 c_v 值，并设定 k 值的范围为 1 至 50。最后，综合不同主题数 k 的一致性度量 c_v 值及其主题分布情况选出 LDA 主题模型对应的最优主题数。模型结果如图 4 所示。

主题模型的一致性指数越高，其分类结果越优[32]。如图 4 所示，当主题数量 k 值为 5 或 9 时，模型的一致性指数取得较高值，同时，通过比对不同 k 值下

的主题分布情况，发现当一致性指数较低时（如 k = 20、34、50），其主题分布呈现出不均匀、且主题大小差异性较大的特点。因此，通过综合分析一致性度量 c_v 值及主题分布情况，本文认为社交媒体新闻数据集下的LDA主题模型最优主题数 k 值为9。

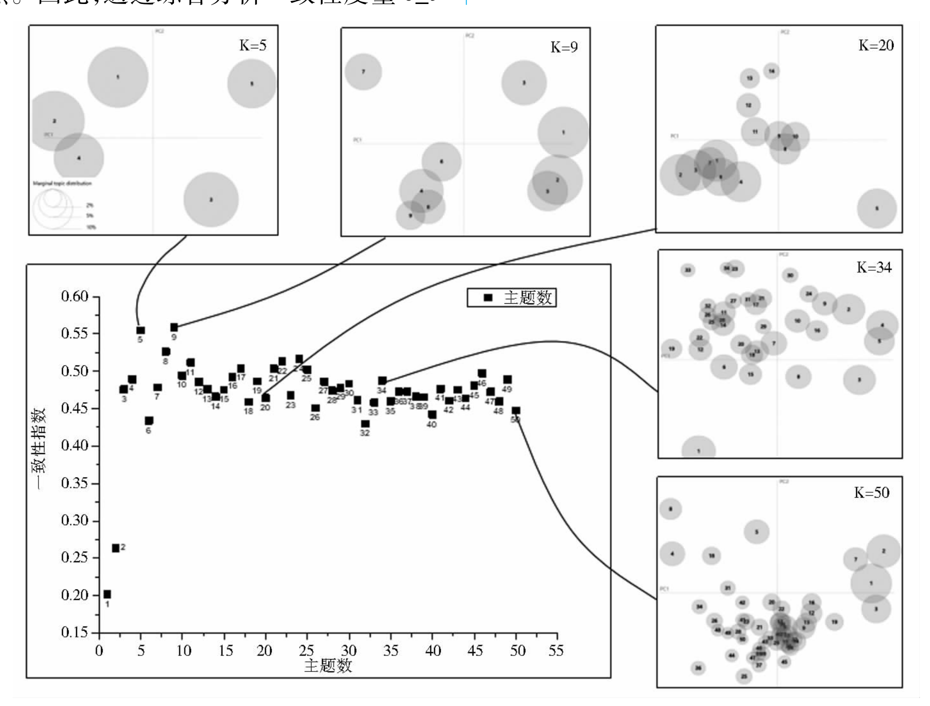


图4　不同主题数下的LDA主题模型结果

4.2　主题过滤结果分析

对基于LDA主题模型提取出的九大主题，通过计算紧密中心度、主题权重以及主题自相关函数三大度量函数评判主题的弱性，过滤出可能包含弱信号的主题。

在本节中，首先计算每个主题与其他主题之间的Hellinger距离，得到一个9＊9的距离矩阵以测量主题的紧密中心度。其次，运用Gensim库衡量每个主题的一致性，并代入权重函数 $W(T)$ 以确定主题的权重。最后基于所有主题每天的文档频率计算主题的自相关函数，其中函数滞后期的确定较为关键。通常，非重叠时间序列的自相关性低于重叠序列的自相关性，且数据越不重叠，其自相关性越低，而大多用于趋势分析的样本之间没有重叠，因此，观测较长滞后期的变化是有益的[38]。

在弱信号检测中，本研究希望最小化主题过滤函数值，即 WK 函数分母部分尽可能大，因此拟通过设置较高的滞后期以减少时间序列之间的重叠周期，使得自相关函数 AC 最小化。因此，选择所观察数据周期的一半作为自相关函数的最佳时滞，即将滞后期定为15。

以上三大度量函数函数的计算都离不开所有主题每天的文档频率，其部分数据如表1所示：

表1　主题文档频率部分数据

日期	T1	T2	T3	T4	T5	T6	T7	T8	T9
2020.11.01	47	35	74	65	43	46	40	36	25
2020.11.02	112	145	86	67	47	58	23	35	64
2020.11.03	189	123	167	172	97	75	45	35	72
2020.11.04	346	263	130	95	53	118	135	74	32
2020.11.05	124	97	114	83	52	46	82	109	34
…	…	…	…	…	…	…	…	…	…
2021.1.09	178	80	98	23	125	84	73	42	64
2021.1.10	93	42	61	63	71	82	42	54	34

图5、图6和图7分别显示了2020年11月、2020年12月和2021年1月的主题过滤结果。图中阴影标记 的是可能包含弱信号的主题过滤结果，这些主题的 WK 函数值高于结果集的1%，而低于结果集的15%。

以月为观测周期，通过主题过滤函数从每月的九大主题中分别提取出T3、T7、T9三个可能包含弱信号的主题，但这些主题内的术语并不都为弱信号，因此本文还将通过术语过滤函数从其中抽取弱信号。

4.3　术语过滤结果分析

LDA主题模型根据每个主题中术语出现的概率对其进行分组和排序。为尽可能地捕获主题内的弱信号，需要从主题中获取足够多的术语。因此，基于主题

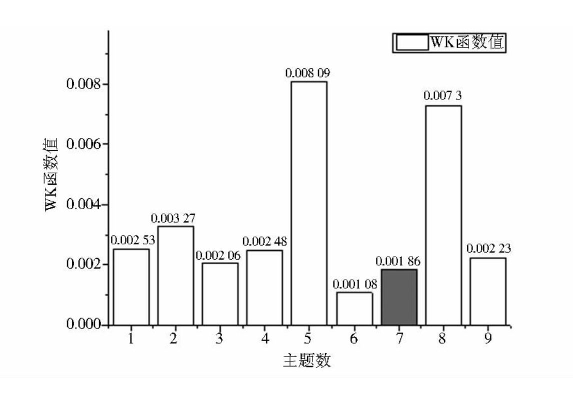


图 5　2020 年 11 月主题过滤结果

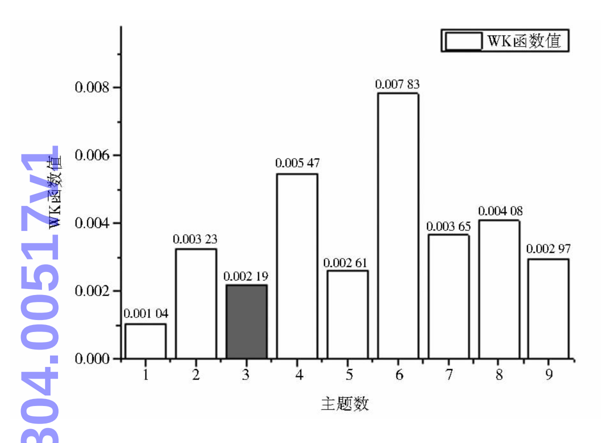


图 6　2020 年 12 月主题过滤结果

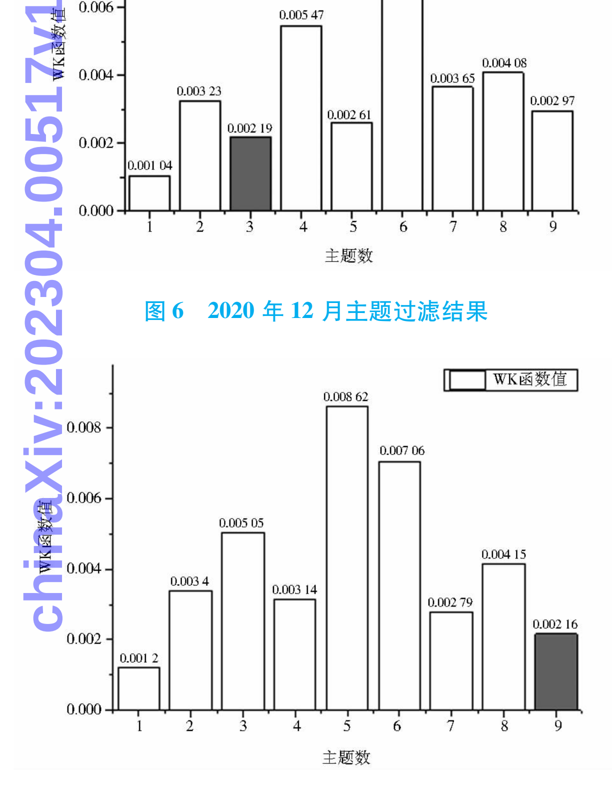


图 7　2021 年 1 月主题过滤结果

过滤结果，本文分别从 2020 年 11 月主题 T7、2020 年 12 月主题 T3 和 2021 年 1 月主题 T9 中提取 500 个术语，并统计每个术语对应的文档频率，运用术语过滤函数从其中提取出弱信号。表 2、表 3 和表 4 分别列出主题 T7、主题 T3 和主题 T9 的弱信号提取结果。

表 2　2020 年 11 月主题 T7 术语过滤结果

主题	术语过滤结果
T7	**持续** 专家 高 实现 建立 国际 勋章 覆盖 **严防** 整改 必要 告诉 严防 运输 污染 百姓 会晤 恐怖主义 协作 医疗 **增加** **恢复** 年底 免费 潜力 **恶化** 升级 **疫苗**

表 3　2020 年 12 月主题 T3 术语过滤结果

主题	术语过滤结果
T3	历史 助力 重大 平等 **温度** **爆发** 坚持 自贸区 通道 资本 全球性 气候变化 旅游 消费 制造业 两岸 循环 **复苏** **感染** **冷冻** 机遇 媒体 安全 绿色 贫困县

表 4　2021 年 1 月主题 T9 术语过滤结果

主题	术语过滤结果
T9	输送 健康 **无症状** 联动 资金 遏制 缓缓 康复 **紧急** 河南 **检测** **持续** 港口 政府 **严重** 投资 海军 典型 增长 崛起 创新 **威胁** 推广 生态 修复

表中部分单词已表现出与疫情的重爆发相关（已加粗），为增强模型的可解释性，运用 BERT 算法对过滤出的术语进行上下文预测，最大化目标单词的概率。

4.4　弱信号提取结果分析

本文欲挖掘与 2021 年 1 月初疫情重爆发的相关弱信号，因此，选用疫情重爆发前 3 月的社交媒体新闻为弱信号提取数据集，尽可能地从社交媒体新闻数据集中获得更大的洞察力。因此，为弥补 LDA 词袋模型的不足，获得更多与上述提取弱信号相关的单词，增强模型结果的准确性、可解释性，本研究使用 BERT 深度学习方法在语义上从上下文对过滤出的术语进行扩展，赋予弱信号更多的情景信息与类似的单词。本研究以能否成功挖掘疫情重爆发的早期预警为弱信号识别模型的效度检验，而表 2 中“持续”一词与研究的内容略相关，对其进行扩展后，发现一些重要的弱信号，如“加剧”“反弹”“恶化”“疾病”等，具体扩展列表如表 5 所示：

表 5　主题 T3 中“持续”术语扩展结果

术语	扩展单词
持续	继续 延绵 不断 产量 **增长** 连续 **继续** 价格 **加剧** **反弹** **恶化** 蔓延 形势 开展 更新 经济 国家 政策 **疾病** 状态 健康 时间 常态化 **转变** 巩固 稳定

结合现实疫情重爆发背景可知，自 2021 年 1 月 10 日起，以河北为首，无症状感染患者急剧增加并逐步向其周边城市扩散，致使我国疫情态势又重新陷入危机状态。在本研究提取的弱信号中，发现了与疫情态势发展相关的预警信号，如在 2020 年 12 月主题 T3 过滤的术语中，“持续”“恶化”“增加”之类的弱信号开始向“爆发”“感染”“复苏”等词转变，而在 2021 年 1 月初主题 T9 过滤出的弱信号中已变为“紧急”“严重”“威胁”等词。《新型冠状病毒感染的肺炎诊疗和防控方案》明确指出新冠病毒惧怕高温，也就是说，新冠病毒怕热不怕冷，气温降低反而导致疫情态势的难以控制，这也是疫情重爆发的关键因素之一。而弱信号提取结

果中“气温”“冷冻”等词，也从气温转冷、病毒传播等方面对疫情重爆发进行了预示。

此外，随着时间的推移，弱信号的演变呈现出以下两种态势：

（1）作为前期弱信号的加强。随着时间的推进，部分弱信号单词在预测程度上会表现出一定的增强，如 2020 年 11 月 T7 中提取出的“疫苗”和 2021 年 1 月 T9 中提取出的“感染”，在背景上确立了与疫情、传染病等具有传播性的疾病息息相关，结合“爆发”“恶化”“复苏”等表现事态发展程度的弱信号，表现出疫情形势不向好的早期预兆，而这一预兆随着时间变化不断增强，2021 年 1 月 T9 中的“严重”“威胁”等信号更是直接警示出事态的严重性。对于具有此类特征的弱信号，决策者要综合考量一段时间的发展趋势，若发现上述与疫情重爆发相关的不向好弱信号，应及时采取应对、防护和控制措施，确保社会人民的安全。而对于正面发展的弱信号如 2020 年 11 月 T7 中的“潜力”“升级”，2020 年 12 月 T3 中的“资本”“机遇”及 2021 年 1 月 T9 中的“崛起”“创新”等，应结合自身发展，在相应的契机下开拓未来市场，把握当前时代下的利好政策。

（2）作为弱信号的类似词。如前所述，弱信号微量、稀有的特点导致提取数量较少，从而增大分析其与未来可能发生情况关联性的难度，运用 BERT 深度学习算法可大程度地丰富检测出的弱信号，如 2020 年 11 月 T7 中的“增加”“严防”等词都扩展出与疫情相关的“疾病”“流行病”等弱信号，警示出疫情重爆发的早期信号。此类弱信号是原数据集提取弱信号的进一步衍生，不再局限于原数据集的限制，而是通过深度学习算法将弱信号之间的关联性进一步发散开来，实现部分弱信号从抽象描述到具体事项的跨越过程。

综合上述，本研究提出的基于 LDA-BERT 融合模型的弱信号识别方法很好地检测出 2020 年 11 月至 2021 年 1 月的社交媒体新闻数据集中的弱信号，同时对其综合进行分析与理解，发掘出弱信号随着时间的推移部分在语义上会逐渐增强的演化特性，在一定程度上为预测出 2021 年 1 月初重爆发的疫情态势提供有益参考。

5 结语

当前，弱信号识别存在诸如人工参与较多，全自动化识别方法的研究尚不完善及模型可解释能力不高等局限性。本研究提出一种基于 LDA-BERT 融合模型的弱信号自动识别系统。运用无监督学习算法 LDA 对预处理后的社交媒体新闻数据集进行主题分类，并提出主题过滤和术语过滤双层过滤函数分别用于从 LDA 主题模型结果中过滤出可能包含弱信号的主题，以及仅从主题中提取可能为弱信号的术语。其中，主题过滤基于三大度量函数函数评判主题的弱性：紧密中心度用 Hellinger 距离衡量主题与主题之间的相似性，主题权重以一致性大小衡量主题的重要性，自相关性在设定的滞后期下观测主题随时间的演变。本研究不接受主题过滤结果中的术语皆为弱信号，因此运用术语过滤函数，以主题内术语的归一化概率和术语对应的文档频率构建模型，仅保留其中潜在的弱信号。最后，为弥补 LDA 词袋模型的不足，增强模型的可解释性，将双层过滤函数的结果输入 BERT 深度学习模型，并输出一系列早期预警信号，可在语义上扩展单词，丰富提取出的弱信号，从上下文中赋予其更多含义。对该模型进行测试，以识别 2021 年 1 月初疫情重爆发相关的弱信号。利用 2020 年 11 月至 2021 年 1 月的社交媒体新闻数据，本文成功检测出如“爆发”“复苏”“恶化”等相关早期预警信号，并归纳总结出弱信号存在的两种态势：一是作为前期弱信号的增强，二是作为弱信号的类似词。同时，以月为周期对提取出的弱信号进行综合分析，发现其随着时间的推移部分在语义上会逐渐增强的演化特性。

本模型解决了当前弱信号识别领域研究人工参与较多、主观性较强的问题，实现了全自动化的弱信号检测过程，大大减少了人类专家的时间和成本。同时提出 LDA-BERT 融合模型及双层过滤函数，在保障仅提取相关弱信号的前提下，充分合理地对弱信号在语义上进行扩展，使模型结果具有较高的解释能力，为情报搜集工作中的弱信号检测提供了新方法、新思路。该方法具有如下优点：①泛化：提取出的弱信号不针对某一特定领域或主题，而是在指定的某段时间内应引起重视的预警信息，决策者可以根据自己的需求选择相关的弱信号。②自动化：弱信号的提取过程中没有人工干预，也不需要关键词的帮助，全自动地对文本进行弱信号检测。③科学化：创新提出双层过滤函数以对主题分类的结果进行过滤，避免了人工筛选的主观性，使其更趋于科学、规范。

此外，本研究仍存在些许不足：①由于弱信号与噪声都具有微量、当前意义不明确、运动缓慢的特点，导致文本去噪工作开展的不够完全；②本研究通过设定较长的滞后期，运用其自相关性能有效地过滤出部分文本噪声，同时也可能过滤出少许有一定价值的弱信

号,不能完全无损地从文本集中对其进行提取。因此,未来将着重研究弱信号识别领域的文本去噪工作,为决策者提供更精准的预警信息。

参考文献:

[1] 吴金红,张飞,鞠秀芳. 大数据:企业竞争情报的机遇、挑战及对策研究[J]. 情报杂志,2013,32(1):5-9.

[2] 邵波, 宋继伟. 反竞争情报预警中的风险识别及排序[J]. 情报理论与实践, 2007, 30(5):642-645.

[3] WISSEMA H. Driving through red lights[J]. Long range planning, 2002, 35(5):521-539.

[4] MUHLROTH C, GROTTKE M. A systematic literature review of mining weak signals and trends for corporate foresight[J]. Journal of business economics, 2018, 88(5): 643-687.

[5] 蒋甜,刘小平,刘会洲. 基于关键词关联度指标(KRI)进行 LDA 噪声主题过滤的方法研究[J]. 图书情报工作,2020,64(3):92-99.

[6] YOON J. Detecting weak signals for long-term business opportunities using text mining of Web news[J]. Expert systems with applications, 2012, 39(16):12543-12550.

[7] COFFMAN B. Weak signal research, part I: introduction[EB/OL]. [2021-07-10]. http://legacy. mgtaylor. com/mgtaylor/jotm/winter97 /jotmwi97. htm.

[8] ROSSEL P. Weak signals as a flexible framing space for enhanced management and decision-making[J]. Technology analysis and strategic management, 2009, 21(3):307-320.

[9] MENDONA S, PINAEC M, KAIVO-OJA J, et al. Wild cards, weak signals and organisational improvisation[J]. Futures, 2004, 36(2): 201-218.

[10] SANDRO M, GUSTAVO C, JOAO C. The strategic strength of weak signal anal-ysis[J]. Futures, 2012, 44(3): 218-228.

[11] IGOR ANSOFF H. Managing strategic surprise by response to weak signals[J]. California management review, 1975, 18(2): 21-33.

[12] HOLOPAINEN M, TOIVONEN M. Weak signals: ansoff today[J]. Futures, 2012, 44(3): 198-205.

[13] 沈固朝. 信号分析:竞争情报研究的又一重要课题[J]. 图书情报工作, 2009, 53(20):11-59.

[14] 单彬. 认知视角下的弱信号分析及实证研究[D]. 北京:中国人民解放军军事医学科学院,2014.

[15] 赵小康. 弱信号:识别、探测与应对[J]. 情报杂志,2010,29(1):159-163.

[16] GRIOL-BARRES I, MILLA S, CEBRIÁN A, et al. Detecting weak signals of the future: a system implementation based on text mining and natural language pro-cessing[J]. Sustainability, 2020, 12(19): 1-22.

[17] GRIOL-BARRES I, MILLA S, MILLET J. System implementation for detection of future weak signals using text mining[J]. Revista española de documentación científica, 2019, 42(2): e234-e234.

[18] 邓胜利,林艳青,王野. 企业竞争弱信号的特征提取与定量识别研究[J]. 图书情报工作,2016,60(10):67-75.

[19] HIRSCHBERG J, MANNING C D. Advances in natural language processing[J]. Science, 2015, 349 (6245):261-266.

[20] YOUNG T, HAZARIKA D, PORIA S, et al, Recent trends in deep learning based natural language processing[J]. Journal of engineering, 2018, 13(3): 55-75.

[21] DIENG A B, RUIZ F J R, BLEI D M. Topic modeling in embedding spaces[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 439-453.

[22] PEPIN L, KUNTZ P, BLANCHARD J, et al. Visual analytics for exploring topic long-term evolution and detecting weak signals in company targeted Tweets[J]. Computers & industrial engineering, 2017, 112(2):450-458.

[23] GUTSCHE T. Automatic weak signal detection and forecasting[D]. Enschede: University of Twente, 2018.

[24] 庄穆妮,李勇,谭旭,等. 基于 BERT-LDA 模型的新冠肺炎疫情网络舆情演化仿真[J]. 系统仿真学报,2021,33(1):24-36.

[25] MAITRE J, MÉNARD M, CHIRON G, et al. A meaningful information extrac-tion system for interactive analysis of documents[C]//2019 international conference on document analysis and recognition. Sydney: IEEE. 2019. 92-99.

[26] LEE K, FILANNINO M, UZUNER Ö. An empirical test of GRUs and deep contextualized word representations on de-identification[J]. Studies in health technology and informatics, 2019, 264(5):218-222.

[27] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of ma-chine learning research, 2003(3): 993-1022.

[28] 赵凯,王鸿源. LDA 最优主题数选取方法研究:以 CNKI 文献为例[J]. 统计与决策,2020,36(16):175-179.

[29] CHANG J, GERRISH S, WANG C, et al. Reading tea leaves: how humans inter-pret topic models[C]//Neural information processing systems. New York: Curran Associates. 2009:288-296.

[30] NEWMAN D, LAU J H, GRIESER K, et al. Automatic evaluation of topic coherence[C]//The 2010 annual conference of the North American chapter of the Association for Computational Linguistics. Los Angeles: Association for Computational Linguistics. 2010:100-108.

[31] 黄佳佳,李鹏伟,彭敏,等. 基于深度学习的主题模型研究[J]. 计算机学报,2020,43(5):827-855.

[32] RODER M, BOTH A, HINNEBURG A. Exploring the space of topic co-herence measures[C]//Proceedings of the eighth ACM international conference on Web search and data mining. New York: Association for Computing Machinery, 2015:399-408.

[33] YOKOYAMA S, SANADA H. Logistic regression model for predicting language change[A]//KOHLER R. Issues in quantitative linguistics. Lüdenscheid: RAM-Verlag, 2009.

[34] THORLEUCHTER D, POEL D. Weak signal identification with semantic Web mining[J]. Expert systems with applications, 2013, 40(12): 4978－4985.

[35] CHUANG J, MANNING C D, HEER J, Termite: visualization techniques for assessing textual topic models[C]//Proceedings of the international working conference on advanced visual interfaces. New York: Association for Computing Machinery, 2012: 74－77.

[36] SIEVERT C, SHIRLEY K. LDAvis: a method for visualizing and interpreting topics[C]//Proceedings of the workshop on interactive language learning, visualization, and interfaces. Baltimore: Association for Computational Linguistics, 2014:63－70.

[37] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C/OL]. NAACL-HLT,2019(1).[2021－05－25]. https://arxiv.org/abs/1810.04805.

[38] FRANKLAND R, SMITH A D, SHARPE J, et al. Calibration of VaR models with overlapping data[J]. British actuarial journal, 2019(24). [2021－06－25]. http://dx. doi. org/10. 1017/S1357321719000151.

[39] EL AKROUCHI M, BENBRAHIM H, KASSOU I. Early warning signs detection in competitive intelligence[C]//The 25th International Business Information Management Association conference. Amsterdam: Association for Computing Machinery, 2015: 512－524.

[40] BLANCO S, LESCA H. Business intelligence: integrating knowledge into selec-tion of early warning signals[EB/OL].[2021－06－25]. http://veille-strat egique. eolas-services. com.

作者贡献说明：
杨波:论文修正与批阅;
邵婉婷:模型构建与论文撰写。

Research on Weak Signal Recognition Based on LDA-BERT Fusion Model

Yang Bo[1,2] Shao Wanting[1,2]

[1] School of Information Management, Jiangxi University of Finance and Economics, Nanchang 330013

[2] Institute of Information Resources Management, Jiangxi University of Finance and Economics, Nanchang 330013

Abstract: [**Purpose/significance**] Aiming at the problem that the existing weak signal automatic recognition research is still incomplete, this paper proposes a weak signal automatic recognition method based on the LDA-BERT fusion model. [**Method/process**] Based on the unsupervised LDA topic model, the text data set was classified by topic, and the topic and term double-layer filter function was constructed to extract early warning signals from the results of topic classification. The weakness of the topic was evaluated by the three major metrics of close centrality, topic weight and topic autocorrelation, and weak signals were extracted based on the normalized frequency and probability of terms within the topic. Finally, the BERT deep learning model was used to expand the weak signal context and similar words from the semantic level. [**Result/conclusion**] Taking the re-eruption of the epidemic in early January 2021 as an example, the constructed system model was verified using the social media news data set of the three months before the outbreak. The experimental results show that the method can effectively detect the relevant weak signals and dig out the evolution characteristics of the weak signals that gradually increase over time. In addition, the fusion model not only realizes the automatic identification of weak signals, but also shows stronger result interpretability than a single model.

Keywords: weak signals LDA-BERT model new crown pneumonia epidemic